Predicting pregnancy status from mid-infrared spectroscopy in dairy cow milk using deep learning

Brand, W; Wells, AT; Smith, SL; Denholm, SJ; Wall, E; Coffey, MP

Published in:
Journal of Dairy Science

DOI:
10.3168/jds.2020-18367

Print publication: 01/04/2021

Document Version
Version created as part of publication process; publisher's layout; not normally made publicly available

J. Dairy Sci. 104
https://doi.org/10.3168/jds.2020-18367

Predicting pregnancy status from mid-infrared spectroscopy in dairy cow milk using deep learning

W. Brand, A. T. Wells, S. L. Smith, S. J. Denholm, E. Wall, and M. P. Coffey*
Scotland's Rural College (SRUC), Peter Wilson Building, Kings Buildings, West Mains Road, Edinburgh, EH9 3JG, UK

ABSTRACT

Accurately identifying pregnancy status is imperative for a profitable dairy enterprise. Mid-infrared (MIR) spectroscopy is routinely used to determine fat and protein concentrations in milk samples. Mid-infrared spectra have successfully been used to predict other economically important traits, including fatty acid content, mineral content, body energy status, lactoferrin, feed intake, and methane emissions. Machine learning has been used in a variety of fields to find patterns in vast quantities of data. This study aims to use deep learning, a sub-branch of machine learning, to establish pregnancy status from routinely collected milk MIR spectral data. Milk spectral data were obtained from National Milk Records (Chippenham, UK), who collect large volumes of data continuously on a monthly basis. Two approaches were followed: using genetic algorithms for feature selection and network design (model 1), and transfer learning with a pretrained DenseNet model (model 2). Feature selection in model 1 showed that the number of wave points in MIR data could be reduced from 1,060 to 196 wave points. The trained model converged after 162 epochs with validation accuracy and loss of 0.89 and 0.18, respectively. Although the accuracy was sufficiently high, the loss (in terms of predicting only 2 labels) was considered too high and suggested that the model would not be robust enough to apply to industry. Model 2 was trained in 2 stages of 100 epochs each with spectral data converted to gray-scale images and resulted in accuracy and loss of 0.97 and 0.08, respectively. Inspection on inference data showed prediction sensitivity of 0.89, specificity of 0.86, and prediction accuracy of 0.88. Results indicate that milk MIR data contains features relating to pregnancy status and the underlying metabolic changes in dairy cows, and such features can be identified by means of deep learning. Prediction equations from trained models can be used to alert farmers of nonviable pregnancies as well as to verify conception dates.

Key words: pregnancy status, deep learning, transfer learning, genetic algorithms

INTRODUCTION

Pregnancy status is an essential phenotype in dairy cattle and important in managing the reproductive—and subsequent production—performance of the herd. Over the course of lactation, the milk yield peaks and then declines; however, poor reproductive performance allows more cows to lactate far after they have reached their peak, thus reducing profitability. To attain optimal herd efficiency, farmers aim for a 365-d calving interval, meaning the cow must be inseminated 80 d postpartum and maintain this pregnancy throughout her lactation. The longer it takes to determine that the cow has not maintained the pregnancy, the greater the financial implications. Cows thought to be pregnant and identified late in lactation as being empty are often culled because a subsequent potential pregnancy does not fit the farms calving pattern or justify the prolonged dry period of the cow. Pregnancy diagnosis is routinely carried out by a veterinarian, usually using rectal palpation, approximately 3 wk after insemination (Sheldon and Noakes, 2002). On establishing pregnancy, the cow is assumed to be in calf unless she begins displaying signs of estrus. The ability and speed with which estrus is detected is dependent on the quality of management and detection aids on farm (Roelofs et al., 2010). Pregnancy diagnosis can also be established from a milk sample by measuring the concentration of progesterone at 24 d, with accuracies of 83.3 and 85%, respectively (Muhammd et al., 2000; Sheldon and Noakes, 2002). Pregnancy has also been shown to affect milk composition (Olori et al., 1997; Penasa et al., 2016; Lainé et al., 2017), and was highlighted via calibration of spectral data from mid-infrared (**MIR**) spectroscopy of milk samples collected as part of routine milk recording (Lainé et al., 2017). This is of particular interest because prediction of pregnancy status from samples collected as part of routine milk recording could provide a faster detection

Received February 14, 2020.
Accepted October 1, 2020.
*Corresponding author: mike.coffey@sruc.ac.uk

method that is noninvasive, cost-effective, and able to be applied on a regular basis.

Infrared radiation is the section of the electromagnetic radiation spectrum with wavelengths longer than light (780 nm^{-1} mm), making it invisible to the human eye. The mid-infrared region of the infrared spectrum is between 3 and 50 μm. When MIR radiation hits an object, the molecules from which it is composed absorb the energy and begin to rotate and vibrate. Rather like a fingerprint, the rotational and vibrational patterns are characteristic of different molecules, allowing identification of molecules by their pattern of absorbance. Mid-infrared spectroscopy is routinely used for the quantification of fat and protein contents of milk samples; however, several other compounds expressed in milk samples could also be identified through MIR spectra and used to monitor the health status of the lactating cow. Already MIR spectra have been calibrated to develop prediction equations for, among others, fatty acid content (Soyeurt et al., 2006; Wojciechowski and Barbano, 2016), mineral content (Toffanin et al., 2015), body energy status (McParland et al., 2011; Smith et al., 2019), lactoferrin (Soyeurt et al., 2012), and methane emissions (Dehareng et al., 2012). Additional studies have focused on pregnancy diagnosis and have shown that signals in the milk MIR can provide an indication of a change in the pregnancy status of cows; however, mixed success in calibrating milk MIR spectra to predict pregnancy status has been reported (Lainé et al., 2014; Toledo-Alvarado et al., 2018; Delhez et al., 2020).

Previous studies looking at phenotype prediction from milk MIR spectra have mostly focused on using partial least squares (**PLS**) analysis to develop prediction equations (see studies mentioned above and review by De Marchi et al., 2014). The volume of data, combined with the computing power, available to scientists today presents new techniques, such as machine learning and artificial neural networks, and opportunities to delve deeper in investigating relationships between MIR spectra and economically important phenotypes.

Artificial neural networks are computer systems inspired by the biological neural networks found in mammalian brains (Ciresan et al., 2011) with extensive networks of interconnected neurons. Deep neural networks are similar to artificial neural networks, except that they include 2 or more hidden layers, which enables them to discover features in complex, high-dimensional data for classification or detection by means of representation-learning methods (LeCun et al., 2015). Advances in deep neural networks have demonstrated the ability to accurately classify complex data from several disciplines, especially for computer vision (J.-H. Jacobsen, E. Oyallon, S. Mallat, and A. W. M. Smeulders, unpublished data, "Multiscale hierarchical convolutional networks"). Deep neural networks are essentially feed-forward systems where information is passed in a single direction. Convolutional neural networks mimic the mammalian brain even further by using supervised back-propagation to update older assumptions with newly acquired knowledge during training, by means of sampling and subsampling maps (Ciresan et al., 2011). These convolutional neural networks are essential to the extraction of high-level features from abstract data to improve the predictability of deep classifier layers. Transfer learning utilizes all the same design requirements but exploits the fact that data from one feature space and distribution can be used to classify data in another feature space and distribution (Pan and Yang, 2010). This means that models can be trained on data sets where training data is excessive and subsequently used on sparse data for further training. Transfer learning models are mostly available for computer vision tasks such as classifying images into discrete categories. Following a machine learning approach, a pilot study by our group confirmed that milk MIR spectra contained features relating to pregnancy status and underlying metabolic changes in dairy cows and that those features could be identified using artificial neural networks (Brand et al., 2018). This work was further extended and applied to milk MIR spectral data to successfully predict bovine tuberculosis status of individual cows (Denholm et al., 2020).

The objective of this study was to use deep learning to model the relationship between milk MIR spectral data and pregnancy status in dairy cows. The ability to determine whether or not a cow is pregnant from her spectral profile alone would provide not only a noninvasive and low-cost method to diagnose pregnancy but also the ability to monitor the pregnancy status of the entire herd throughout lactation. More importantly, it would enable the farmer to be alerted to any changes in status between recordings, such as confirmation of pregnancy following insemination (i.e., moving from a not-pregnant state to a pregnant state) as well as loss of pregnancy (i.e., moving from a pregnant state to a not-pregnant state).

MATERIALS AND METHODS

Acquisition and Scope of Data

Mid-infrared analysis of milk samples was carried out by National Milk Records (Chippenham, UK) using Foss spectrometers (Foss Electric A/S, Hillerød, Denmark), based at the National Milk Laboratories

(Glasgow, UK). Data were collected as part of routine milk recording services in the United Kingdom and electronically transferred to Scotland's Rural College (Edinburgh, UK) nightly on a continuous basis. Sampling intervals were 30 d on average.

The process of selecting records for analysis was based on our perceived ability to classify cows as pregnant or nonpregnant. The only certain way is to use records from cows that have calved again and assume that before the calculated or recorded insemination the cow was not pregnant, and afterward she was pregnant. Insemination records are not always recorded; thus all data after the "last" recorded insemination could not be assumed pregnant—the farmer could, as is often the case, stop recording inseminations when the cow is not seen bulling and subsequently start recording again some time later. Between recording periods, even with a confirmed pregnancy, the time when a cow was pregnant and then was not is too imprecise. To avoid introducing such uncertain and imprecise records into our training set, they were excluded. Thus, milk MIR spectral records from animals after parturition and before their first insemination were labeled as nonpregnant for the training data set. Records between the last insemination and the subsequent calving with a gestation length between 240 and 284 d were labeled as pregnant records for the data set.

The amount of records for confirmed nonpregnant animals was the limiting factor, as the distribution of animals in both categories in the training set should be close to equal (LeCun et al., 2015). After labeling the data, a total of 3 million spectral records from 697,671 animals, born between 1999 and 2016, were available for further analysis.

Pretreatment and Standardization of Mid-Infrared Data

The MIR spectrum is stored as 1,060 data points spanning 900 to 5,000 cm^{-1}; each point represents the pattern of absorption of infrared light at a given wavelength (Grelet et al., 2015). Spectral data were converted from transmittance to absorbance using a $\log_{10}{}^{-0.5}$ transformation. Additionally, to account for the difference between different MIR instruments, the data were standardized in accordance with the protocol set out by the EU-funded OptiMIR Project (Friedrichs et al., 2015). Standardization files are received routinely from the Centre Walloon de Recherches Agronomiques (Gembloux, Belgium). This ensures that comparisons can be made across any tools developed within the same dairy network or results collected where this standardization has been applied.

Model Development

Two models were developed (labeled model 1 and model 2) and investigated by applying different deep learning techniques. The development of model 1 involved a multistep approach and used genetic algorithms (**GA**) to reduce the dimensionality of the MIR spectra by eliminating wave points that were not significant to predicting pregnancy status (feature selection). Genetic algorithms are computer programs that evolve in ways that resemble natural selection to solve complex problems. All GA were implemented on a representative subset of 100,000 records from the MIR data. The purpose of the GA was not to predict pregnancy, but rather to investigate the possibility of using a smaller subset of MIR wave points when predicting pregnancy as well as defining an appropriate deep neural network that can predict pregnancy status from a subset of MIR wave points. Each wave point was randomly assigned a discrete weighting of 0 or 1 that determined whether a wave point would be selected into the feature space (GA1) for feature selection. A visual description of how this was implemented can be seen in Figure 1. The first generation consisted of 50 individuals, each holding a random set of wave points for selection. A control test was performed on all 1,060 wave points to benchmark the predictive difference between using all wave points and using a subset. Each individual was evaluated on accuracy using a k-nearest neighbors approach. Individuals with the highest accuracy were subsequently selected to generate new individuals for future generations or iterations. Accuracy was defined by Equation [1] as follows:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN), \quad [1]$$

where TP, TN, FP, and FN represent total numbers of true positives, true negatives, false positives, and false negatives as predicted by the model, respectively.

The "fittest" individual after 250 iterations was selected, and its selected wave points were used for the next GA. A second set of GA were implemented on the selected wave points from GA1 by assigning continuous weighting factors between 0 and 1 to ensure that the subset of wave points was indeed still trainable (GA2) for feature extraction. The benchmark in GA2 was the selected individual from GA1. This step was performed to ensure that the reduced feature set could still be subjected to training and that too many features were not eliminated. In both GA1 and GA2, the prediction accuracies for the "fittest individual" and the population average (average accuracy of all individuals in a generation) were logged at the end of iteration. A third GA was trained with the reduced feature set

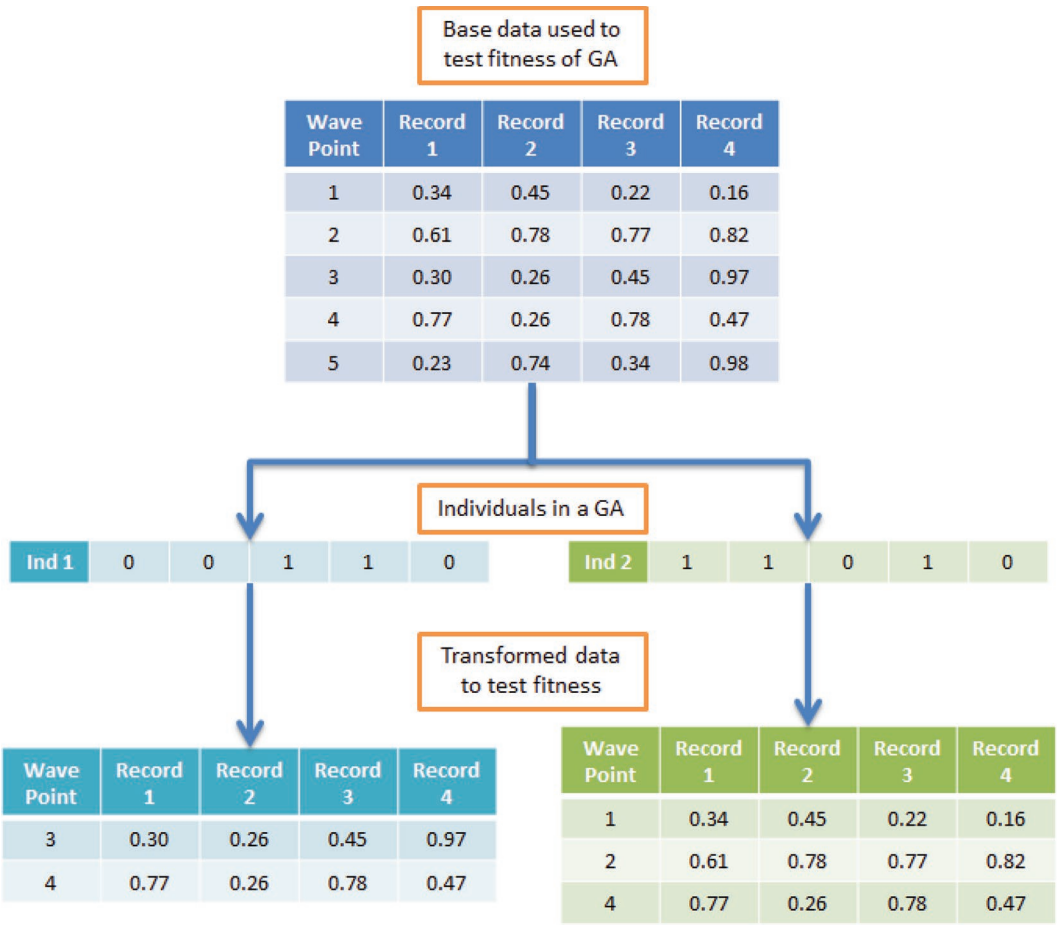

Figure 1. Visual interpretation of how data are transformed through different components in a genetic algorithm when applied for feature selection. GA = genetic algorithm; Ind = individual; Wave point = location within the mid-infrared spectrum of 1,060 wave points.

to design an optimum deep neural network by setting each individual in the base population as a random network configuration and evolved for several generations (GA3). The resulting neural network architecture was subsequently applied to the spectral data for further training and optimization on a larger data set of 3,000,000 spectral records, evenly distributed between pregnant and nonpregnant.

The development of model 2 involved obtaining a pretrained model called DenseNet (Huang et al., 2017) and adapting it to MIR spectra classification through transfer learning. Mid-infrared spectra records were individually converted into gray-scale images with dimensions of 53 × 20 pixels (from the original 1,060 wave points). Pretrained models such as DenseNet are trained on millions of images and are well adapted to extracting high-level features from abstract data, such as images from its deeper convolutional layers. This allows for more robust models in subsequent training, and a smaller data set can be used. Model 2 was trained on only 10,000 spectral images, equally distributed between both labels and spanning different stages of lactation. The model was trained for 100 epochs and with the convolutional layers set to non-trainable. This allowed the model to understand the MIR images first by training the dense, classifier layers only. Subsequently the model was trained for another 100 epochs and with convolutional layers available for training with a small learning rate. An inference data set of 1,000 spectral images was used to test the quality of predictions from model 2.

Deep learning models are typically trained on a data set split into 2 subsets of data, one for training and learning (the training set) and a second for validating during training (the validation set). Both of these data sets are passed to the model during training with the features (MIR spectral wavelengths) as well as the labels (binary pregnancy status). The ratio used to split a data set into training and validation sets is usually 4:1 for training and validation, respectively; this ratio was maintained when creating training and validation sets in the present study. Models were evaluated on 2

metrics: accuracy, as defined in Equation [1], and loss, obtained via a loss function. In the case of categorical labels, such as in the present study, a Softmax activation function (Equation [2]) is applied to the final output layer of the network before applying a suitable loss function: here, categorical cross-entropy (Equation [3]). Note that the Softmax activation function normalizes the output of the network in the range (0, 1), thus providing a discrete probability distribution, such that the components of the resulting output vector sums to 1:

$$s(x_i) = \frac{e^{x_i}}{\sum_{j}^{C} e^{x_j}}; \quad [2]$$

$$Loss = -\sum_{i}^{C} t_i \log\left[f(s)_i\right]. \quad [3]$$

Here, x is the observation from $j = 1$ to C; C is the number of possible class labels (in this case C is 2, representing each pregnancy status); e is the standard exponential function; t is the target vector; and $f(s)$ is the Softmax probability obtained by applying Equation [2].

Loss (Equation [3]) helps us to interpret the confidence of the model's predictions and can range from zero to infinity, the former being the ideal goal. Although the accuracy of prediction for binary labels can be high, so too can the loss; thus, optimizing (i.e., reducing) the loss metric close to zero ensures that the model is robust in its predictions.

Three further metrics commonly used in machine and deep learning were also calculated for resultant models to determine performance. These included precision, recall, and F_1-score. Precision (the positive predictive value) was calculated via Equation [4] and represents the proportion of positive predictions that were verified as correct. Recall (sensitivity, or true positive rate) was calculated via Equation [5] and represents the proportion of true positives the model identified correctly. Finally, the F_1-score (used in the analysis of binary labels) was calculated via Equation [6] and represents the harmonic mean of precision and recall.

$$Precision = TP/(TP + FP), \quad [4]$$

$$Recall = TP/(TP + FN), \quad [5]$$

and

$$F_1\text{-}score = 2 \times (Precision \times Recall)/(Precision + Recall), \quad [6]$$

where TP, FP, and FN are numbers of true positives, false positives, and false negatives as predicted by the model, respectively.

Both models were developed and trained on a NVIDIA DGX Station with 4 NVIDIA Tesla V100 GPU cards (NVIDIA Corporation, Santa Clara, CA). This greatly improved training time, especially as the second training of model 2 had 28,744,386 trainable parameters that require updating at each epoch. The open-source TensorFlow application programming interface from Google Inc. (Abadi et al., 2015) was used to develop model 1, and the Fast.ai application programming interface (Howard and Gugger, 2020) was used to develop model 2.

Partial least squares analysis has been the technique generally used to date when predicting phenotypes from milk MIR spectra (see review by De Marchi et al., 2014). When the response variable is categorical, a PLS variant known as partial least squares discriminant analysis (**PLS-DA**) can be applied, as has been shown previously, to develop prediction models of pregnancy status from milk MIR spectra (Delhez et al., 2020). Therefore, in addition to the 2 deep learning models previously described, a PLS-DA was also applied to a subset of the data (balanced for label). Before applying the PLS-DA, data were smoothed to remove baseline variation by calculating the first derivative of the raw spectra—that is, subtracting from each wavelength value the immediately preceding wavelength value (McParland et al., 2011; Soyeurt et al., 2011; Smith et al., 2019). Partial least squares discriminant analysis was then carried out using Python 3.5 (van Rossum, 1995) and the Scikit-learn machine learning package (Pedregosa et al., 2011). Cross-validation (random, 10-fold CV) was used to evaluate PLS-DA model performance and enable comparison between the different models.

RESULTS

Model 1

The configurations for all GA are summarized in Table 1. After each generation, all individuals in the population were evaluated for fitness, based on the model's ability to accurately predict pregnancy status from its features, and subsequently ranked by accuracy in descending order. The first 40% were selected as parents for the next generation. The rest of the population were individually given a 10% chance of being randomly selected as parents as well, to maintain variation. These individuals would then be randomly paired to create new individuals until the population capacity was reached. The process would then repeat itself until

Table 1. Standard configuration of genetic algorithms used for feature selection and neural network architecture

Option	Factor	Comment
Retention rate	0.4	Proportion of individuals selected as parents for next generation, order by best to worst on accuracy
Random selection rate	0.1	Random chance of nonselected individuals to be selected as a parent for next generation
Mutation rate	0.2	Random chance that an individual will be modified for a specific feature

250 iterations were completed, which was sufficient to allow for favorable random mutations in the equations.

The first genetic algorithm (GA1) selected 196 features after 157 iterations. Using all 1,060 MIR wave points could predict pregnancy status with an accuracy of 0.8225. The reduced feature set received an accuracy of 0.8501, and the average of the population was 0.8477. An increase of more than 2% in accuracy, using 18.49% of the original feature set, was favorable and less computationally demanding. Concerns about whether too much information had been removed were addressed with GA2, which showed that the fittest individual could be further trained to an accuracy of 0.8731, and the 196 wave points were used in GA3 to obtain optimum neural network architecture. The third genetic algorithm (GA3) suggested a convolutional deep neural network with a Softmax activation function (Equation [2]). The Softmax activation function is a normalized exponential function for multiclass classification and was applied to the output layer of the classifier.

Subsequent training of the neural network on the full data set of 3 million records and 196 features converged after 162 epochs. The validation accuracy and loss are summarized in Figure 2. The training accuracy reached its peak at step 227,413 with a value of 0.90. Despite this, the model reached its lowest loss at step 729,142 with a value of 0.18 and an accuracy of 0.89. Model 1 was not considered for further evaluation and inference due to the relatively high loss of training, although it was noted that the accuracy achieved was higher than with the *k*-nearest neighbors algorithm.

Model 2

The training accuracy and losses of model 2 for each epoch are summarized in Figure 3. Accuracy improved rapidly from the start of training until epoch 33, to 0.925, and thereafter increased at a lower rate to epoch 100 (0.955). The second phase of training showed an initial deterioration of accuracy, but this improved by epoch 157, and subsequently the accuracy converged to 0.9725. Similarly, the losses showed rapid improvement from start of training, followed by a gradual improvement for the first phase of training. Training loss and validation converged at 0.057909 and 0.080359, respectively.

A confusion matrix of the inference data set of model 2 is shown in Table 2. Overall accuracy of prediction was 0.877 with a recall (sensitivity) of 0.894 and precision (positive predictive value) of 0.8646. Recall is disproportional to false negative rate and showed that the model had a low incidence of falsely predicting nonpregnant animals. The F_1 score (harmonic mean) was 0.8791 and corresponded well with the overall accuracy of the test.

PLS-DA Model

Results from the PLS-DA are summarized in Table 3. Overall accuracy of the cross-validation was 0.77, with a recall, precision, and F_1 score of 0.73, 0.80, and 0.76, respectively. Specificity was relatively high (0.82), and, again, overall accuracy and F_1 score corresponded well.

DISCUSSION

The GA proved to be an efficient technique in identifying features in MIR spectral data. The 196 wave points selected by the GA aligned with the wave points selected from the OptiMIR Project (Friedrichs et al., 2015). The GA proved to be versatile in their applications and were easily interpreted. Ultimately, model 1 was not considered appropriate for further interrogation due to its higher loss metric. Convolutional neural networks are widely used for classifying images (Yim et al., 2015) and use padding as a subsampling tool (R. K. Srivastava, K. Greff, and J. Schmidhuber, unpublished data, "Training very deep networks," https://arxiv.org/abs/1507.06228) to remove background noise from the edges of images. Zero-padding was specifically not used in the architecture of the convolutional neural networks, because the 196 features were already subsampled and it was imperative that feature detection occurred on the edges of the convolutional layers.

Influence of Stage of Lactation

Training records classified as not-pregnant were records obtained before first insemination and, therefore, early in lactation, as opposed to pregnant records, which were generally later in lactation. Initial concerns that stage of lactation was being predicted instead of

pregnancy status were not substantiated, whereas examining the predictions as predicted onset of pregnancy varied substantially in the results, and no linear trend could be found. In a previous trial, DIM was fitted as an additional feature, and training accuracies were above 0.97. The model was able to predict pregnancy status with high accuracy, based solely on stage of lactation, and could not identify a single record where a pregnancy was terminated during the lactation. An almost linear increase in the probability of pregnancy was observed as DIM increased. It was concluded that stage in lactation could rather be used to adjust the labels, instead of being used as a feature, by possibly labeling the data as nonpregnant, early pregnant, or late pregnant; as such, DIM was not fitted or made available to the models developed in the present study.

Advantage of Transfer Learning

Transfer learning has the advantage that a robust model for a specific target domain can be obtained by transferring knowledge contained in a different, but related, source domain (Zhuang et al., 2019). By default, this implies that less training data is required to achieve the target model. Model 2 was relatively easy to train with transfer learning, as no prior configuration or investigation on network design was required. Training on spectral images was efficient and faster than parsing text files and converting data types as with model 1. The results showed the capability of the DenseNet model to extract and engineer high-level features from the MIR images. Figure 3 shows no indication of over-fitting (where the model is optimized to predict the validation data set only), which is common in data sets with high complexity (B. Ghojogh and M. Crowley, unpublished data, "The theory behind overfitting, cross validation, regularization, bagging, and boosting: Tutorial"). On the deterioration of accuracy and loss immediately after 100 epochs in model 2, the training of the deep convolutional layers started from epoch 101 and showed that the assigned learning rate was not optimal. Several learning rates were trialed, but all showed a sudden decay of accuracy and loss. A smoother transition may have resulted in an improved model, but these "golden" learning rates could not be obtained, and the best learning rate was found between $1e^{-4}$ and $1e^{-6}$ for phase 2 of training. These learning rates are one of the most important hyperparameters

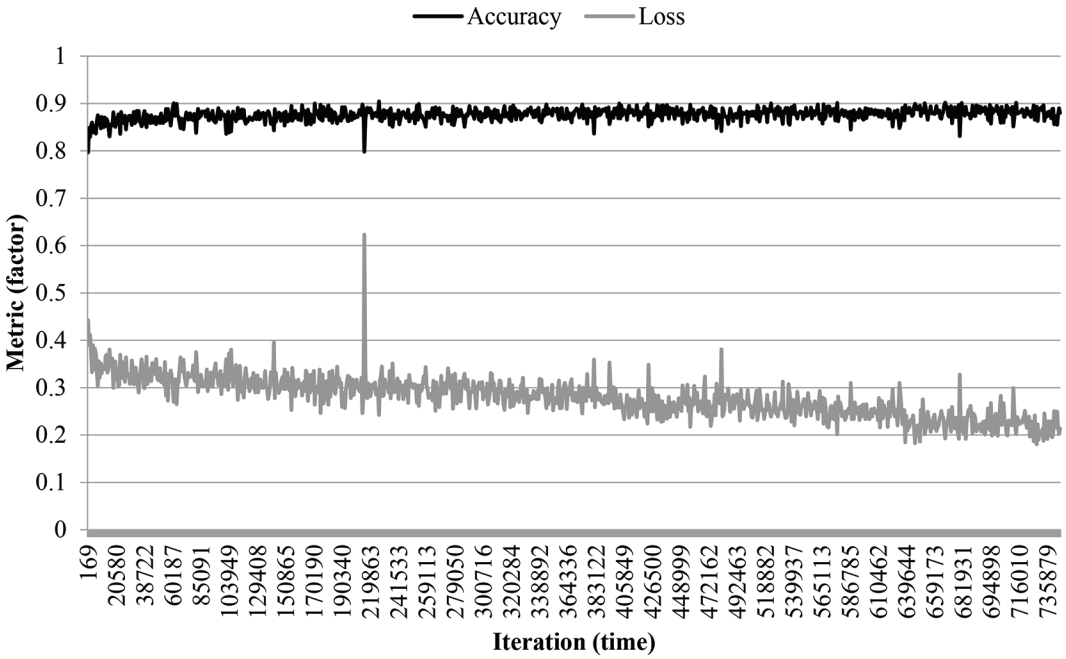

Figure 2. Plot of validation accuracy and loss during training of model 1. Accuracy = (TP + TN)/(TP + TN + FP + FN); $Loss = -\sum_{i}^{C} t_i \log\left[f(s)_i\right]$; where TP, TN, FP, and FN represent total numbers of true positive, true negative, false positive, and false negative predictions, respectively, and $s(x_i) = \frac{e^{x_i}}{\sum_{j}^{C} e^{x_j}}$; where x is the observation from $j = 1$ to C; C is the number of possible class labels (in this case C is 2, representing each pregnancy status); e is the standard exponential function; t is the target vector; and $f(s)$ is the Softmax probability.

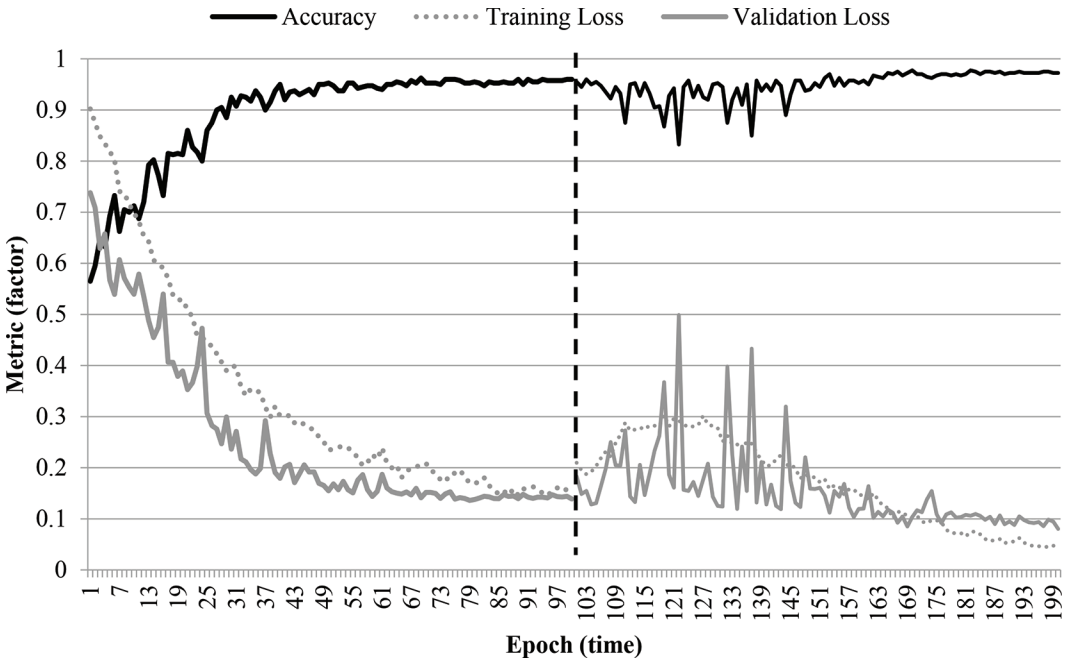

Figure 3. Plot of training metrics for model 2 for accuracy, training loss, and validation loss. The dashed vertical line distinguishes the second phase of training from the first. Accuracy = (TP + TN)/(TP + TN + FP + FN); $Loss = -\sum_{i}^{C} t_i \log\left[f(s)_i\right]$; where TP, TN, FP, and FN represent total numbers of true positive, true negative, false positive, and false negative predictions, respectively, and $s(x_i) = \frac{e^{x_i}}{\sum_{j}^{C} e^{x_j}}$; where x is the observation from $j = 1$ to C; C is the number of possible class labels (in this case C is 2, representing each pregnancy status); e is the standard exponential function; t is the target vector; and $f(s)$ is the Softmax probability.

for a neural network but are network- and data-specific (Howard and Gugger, 2020).

Table 2 shows that 12.3% of the predictions in the inference data set were predicted wrong (false positives and false negatives). Accuracy of predictions can be misleading, because it is discontinuous, especially in the case of binary classification. For example, consider a binary prediction with Softmax activation (Equation [2]) of 0.49 and 0.51 for labels 0 and 1, respectively. If the actual record has a label of 1, the prediction would be 100% correct, and if the label was 0, the prediction would be 100% incorrect. It is, however, clear from the Softmax prediction that the probabilities of both labels are almost equal. From Table 2, the average probabilities of true positive and true negative predictions were 0.971 and 0.968, respectively. In contrast, the average probabilities of false positive and false negative predictions were 0.898 and 0.892, respectively. This suggests that a further distinction can be made in practice by considering predictions with probabilities lower than

Table 2. Model 2 performance: precision, recall, and F_1-scores from inference using model 2[1]

Item	Precision	Recall	F_1-score	Records
Not pregnant	0.86	0.89	0.87	500
Pregnant	0.86	0.89	0.87	500
Accuracy			0.88	1,000

[1]Precision (i.e., positive predictive value) = TP/(TP + FP). Recall (i.e., sensitivity) = TP/(TP + FN). F_1-score = 2 × (precision × recall)/(precision + recall). Accuracy = (TP + TN)/(TP + TN + FP + FN). TP, TN, FP, and FN represent total numbers of true positive, true negative, false positive, and false negative predictions, respectively.

Table 3. Partial least squares discriminant analysis (PLS-DA) model performance: precision, recall, and F_1-scores from 10-fold cross-validation of the PLS-DA model[1]

Item	Precision	Recall	F_1-score	Records
Not pregnant	0.75	0.82	0.78	10,000
Pregnant	0.80	0.73	0.76	10,000
Accuracy			0.77	20,000

[1]Precision (i.e., positive predictive value) = TP/(TP + FP). Recall (i.e., sensitivity) = TP/(TP + FN). F_1-score = 2 × (precision × recall)/(precision + recall). Accuracy = (TP + TN)/(TP + TN + FP + FN). TP, TN, FP, and FN represent total numbers of true positive, true negative, false positive, and false negative predictions, respectively.

0.95 as inconclusive. Table 4 is the confusion matrix of only "conclusive" predictions. The accuracy of predictions improves from 0.877 to 0.9125, and the F_1 score changes accordingly to 0.9142. Sensitivity and specificity of 0.91 and 0.92 are obtained from these results. Results found in literature from pregnancy-associated glycoprotein in dairy cows ranged from 0.96 to 0.99 for sensitivity and 0.87 to 0.95 for specificity (Commun et al., 2016; Dufour et al., 2017; Shephard and Morton, 2018). A point of concern is that 166 predictions were considered inconclusive when applying a minimum threshold for probability.

Table 4. Model 2 performance: precision, recall, and F_1-scores from inference using model 2 when considering predictions with probabilities over 0.95[1]

Item	Precision	Recall	F_1-score	Records
Not pregnant	0.90	0.92	0.91	405
Pregnant	0.92	0.90	0.91	429
Accuracy			0.91	834

[1]Precision (i.e., positive predictive value) = TP/(TP + FP). Recall (i.e., sensitivity) = TP/(TP + FN). F_1-score = 2 × (precision × recall)/(precision + recall). Accuracy = (TP + TN)/(TP + TN + FP + FN). TP, TN, FP, and FN represent total numbers of true positive, true negative, false positive, and false negative predictions, respectively.

Comparison with Previous Studies

Our study is not the first to investigate the utility of using milk MIR spectra in attempting to diagnose pregnancy in dairy cows, but we believe it is the first to attempt to do so using deep learning. As highlighted in our Introduction, previous studies have attempted to calibrate milk MIR spectra to predict pregnancy status in dairy cows, reporting accuracies of 0.90 (Lainé et al., 2014; based on sensitivity and specificity); 0.60 (Toledo-Alvarado et al., 2018; based on area under the receiver operator curve); and, more recently, 0.65 to 0.76 (Delhez et al., 2020; based on area under the receiver operator curve). Prediction equations from these studies were developed using both residual- (Lainé et al., 2014; Delhez et al., 2020) and whole-spectrum MIR profiles (Toledo-Alvarado et al., 2018). Each of these studies highlighted the potential of milk MIR spectra as a predicator of pregnancy status.

Lainé et al. (2014), using a discriminate analysis approach, were able to successfully discriminate between residual spectra from pregnant and nonpregnant cows with a sensitivity of 99.7% and specificity of 86.2% during cross-validation. Residual spectra were generated by subtracting expected open spectra (obtained via a mixed model) from observed spectra. Accuracy was reported to drop significantly (up to 50%) during external validation (Delhez et al., 2020), and an error rate of 55.5% was observed when applied to raw spectra (Lainé et al., 2014).

Toledo-Alvarado et al. (2018), using whole-spectrum MIR from multiple breeds, predicted pregnancy status via generalized linear models fitting a combination of effects (DIM, parity, herd year) in addition to spectra, as well as from milk components. The best accuracies were obtained (area under curve) when herd and year were included with the spectra; lowest prediction accuracy was observed in Holsteins (0.61).

Delhez et al. (2020) adopted a PLS-DA approach and investigated 3 different strategies to discriminate between pregnant and nonpregnant cows based on (1) a single spectrum after insemination, similar to Toledo-Alvarado et al. (2018), but with the addition of including cows with no calving records; (2) residual spectra, similar to Lainé et al. (2014), but using only observed spectra (not modeled); and (3) grouping records by period after insemination. Delhez et al. (2020) reported accuracies (area under curve) of 0.63 and 0.65 for training and testing, respectively, for strategy 1 (with corresponding sensitivity and specificity during testing of 0.65 and 0.56, respectively). For strategy 2, results were similar during testing, with accuracy, sensitivity, and specificity of 0.58, 0.59, and 0.52, respectively. The third strategy observed promising results for records more than 151 d after insemination, reporting average accuracy, sensitivity, and specificity of 0.76, 0.73, and 0.64, respectively.

We observed significantly higher prediction accuracies than the studies highlighted: 88% increasing to 91% when considering only predictions with a confidence over 0.95. This is especially the case when not considering previous results from residual spectra (only observed spectra were used in the development of our models). These higher accuracies may be attributed to a combination factors, including our use of a deep learning approach, phenotype definition, and volume of available data. Moreover, the results obtained by applying a PLS-DA to our data achieved accuracies similar to those obtained by the earlier studies previously discussed; we observed accuracy, sensitivity, and specificity of 0.77, 0.73, and 0.82, respectively, compared with the accuracy, sensitivity, and specificity of 0.76, 0.73, and 0.64, respectively, obtained by Delhez et al. (2020). Additionally, when comparing the PLS-DA method with the DL method used in the development of model 2, not only did we achieve higher accuracies across all metrics calculated using DL (0.91 compared with 0.77) but the development time was also vastly reduced—especially when considering that the data used in the PLS-DA was a (random, balanced) subset of that used to train models 1 and 2.

Use of deep learning in the agricultural space has been limited to date (Howard, 2018), and as such has been met with reservation and suspicion—rightly so without solid proof of validation and evidence of application. Our results highlight high accuracy, sensitivity, and specificity during training, validation, and testing. The training of DL networks involves a methodology similar to a combination of k-fold cross-validation and external validation. After each iteration of the training data (i.e., calibration), the resulting model is then applied to a set of validation data, with results used to update the weights and biases at each node in the network, optimizing the model. The final optimized model is then further applied to an external test data set; the test set is independent of the training and validation sets and simulates a live prediction scenario. Thus we believe that this method of train-validate-test provides a robust indication of model performance.

Definition of the pregnancy status phenotype is an extremely important (if not the most important) aspect of MIR-based prediction. Good-quality and clean phenotypes are not only a crucial requirement of deep learning models (i.e., the labels) but also a crucial requirement of any predictive modeling. In each of the 3 previous studies, and in our own study, the way in which pregnant and nonpregnant (or open) cows are defined differs. It is our belief that by defining nonpregnant records as those between parturition and first insemination we can say with 100% certainty that such records are representative of the nonpregnant class; similarly, for pregnant records (as those between the last insemination and the subsequent calving with a gestation length between 240 and 284 d). This gives us a robust phenotype to pass to the deep learning network.

Finally, it is worth noting the differences in data volume available to each of the previous studies compared with our own. Previously developed models by Lainé et al. (2014), Toledo-Alvarado et al. (2018), and Delhez et al. (2020) used spectra from 68,998, 69,821, and 8,064 cows, respectively; the present study had access to UK national data from 697,671 cows obtained via monthly milk recording over an 8-year period. Moreover, the application of transfer learning greatly reduced the amount of data required to train models, enabling us to create a training data set containing equal numbers of the most accurate phenotypes possible. This, combined with testing on (random) unseen data from throughout the lactation (results in Tables 2 and 3), appears to give a good indicator of pregnancy status. A final test of the models' ability to discriminate pregnant from nonpregnant cows will be obtained through live field testing.

CONCLUSIONS

Deep learning has been shown to be a viable tool in understanding complex data and generation of predictions in new data sets. We believe the present study to be the first to successfully predict pregnancy status (with high accuracy) of dairy cows from observed milk MIR spectral data using a deep learning approach. Convolutional neural networks were found to be an appropriate network architecture to predict pregnancy status from MIR spectra and allowed greater subsampling of features (model 1). Transfer learning proved a viable option for creating high-quality models ready for industry application (91% accuracy during testing). Prediction equations from model 2 can be applied by industry as part of routine milk recording, as a cost-effective monitoring tool to identify possible errors in data recording practices, to verify conception dates, and to alert farmers of nonviable or lost pregnancies as early as possible. Such a tool would also provide an effective enabling service, allowing the farmer to take ownership of the health and fertility of their herd. Finally, such extra information can be generated with no need for additional input or labor on behalf of the farmer or any changes in herd management, and importantly, is noninvasive to the cow.

ACKNOWLEDGMENTS

The authors gratefully acknowledge collaboration with National Milk Records (Chippenham, UK), especially Martin Busfield and Andy Warne. Ian Archibald (Scotland's Rural College, Edinburgh) is acknowledged for curating and managing MIR spectral databases. SJD is funded by the Biotechnology and Biological Sciences Research Council (BBSRC, Swindon, UK; grant no. BB/S009396/1). The authors have not stated any conflicts of interest.

REFERENCES

Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, et al. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. https://www.tensorflow.org/.

Brand, W., A. T. Wells, and M. P. Coffey. 2018. Predicting pregnancy status from mid-infrared spectroscopy in dairy cow milk using deep learning. Page 347 in Abstracts of the 2018 Annual Meeting of the American Dairy Science Association, Knoxville, TN. ADSA, Champaign, IL.

Ciresan, D., U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber. 2011. Flexible, high performance convolutional neural networks for image classification. Pages 1237–1242 in International Joint Conference on Artificial Intelligence (IJCAI), Barcelona, Catalonia, Spain. AAAI Press/International Joint Conferences on Artificial Intelligence, Menlo Park, CA.

Commun, L., K. Velek, J. B. Barbry, S. Pun, A. Rice, A. Mestek, C. Egli, and S. Leterme. 2016. Detection of pregnancy-associated glycoproteins in milk and blood as a test for early pregnancy in dairy

cows. J. Vet. Diagn. Invest. 28:207–213. https://doi.org/10.1177/1040638716632815.

De Marchi, M., V. Toffanin, M. Cassandro, and M. Penasa. 2014. Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits. J. Dairy Sci. 97:1171–1186. https://doi.org/10.3168/jds.2013-6799.

Dehareng, F., C. Delfosse, E. Froidmont, H. Soyeurt, C. Martin, N. Gengler, A. Vanlierde, and P. Dardenne. 2012. Potential use of milk mid-infrared spectra to predict individual methane emission of dairy cows. Animal 6:1694–1701. https://doi.org/10.1017/S1751731112000456.

Delhez, P., P. N. Ho, N. Gengler, H. Soyeurt, and J. E. Pryce. 2020. Diagnosing the pregnancy status of dairy cows: How useful is milk mid-infrared spectroscopy? J. Dairy Sci. 103:3264–3274. https://doi.org/10.3168/jds.2019-17473.

Denholm, S. J., W. Brand, A. P. Mitchell, A. T. Wells, T. Krzyzelewski, S. L. Smith, E. Wall, and M. P. Coffey. 2020. Predicting bovine tuberculosis status of dairy cows from mid-infrared spectral data of milk using deep learning. J. Dairy Sci. 103:9355–9367. https://doi.org/10.3168/jds.2020-18328.

Dufour, S., J. Durocher, J. Dubuc, N. Dendukuri, S. Hassan, and S. Buczinski. 2017. Bayesian estimation of sensitivity and specificity of a milk pregnancy-associated glycoprotein-based ELISA and of transrectal ultrasonographic exam for diagnosis of pregnancy at 28–45 days following breeding in dairy cows. Prev. Vet. Med. 140:122–133. https://doi.org/10.1016/j.prevetmed.2017.03.008.

Friedrichs, P., C. Bastin, F. Dehareng, B. Wickham, and X. Massart. 2015. Final OptiMIR Scientific and Expert Meeting: From milk analysis to advisory tools, Palais des Congrès, Namur, Belgium. Pages 97–124 in Biotechnology, Agronomy, Society and Environment. Presses Agronomiques de Gembloux, Namur, Belgium.

Grelet, C., J. A. Fernández Pierna, P. Dardenne, V. Baeten, and F. Dehareng. 2015. Standardization of milk mid-infrared spectra from a European dairy network. J. Dairy Sci. 98:2150–2160. https://doi.org/10.3168/jds.2014-8764.

Howard, J. 2018. Deep Learning: The tech that's changing everything, except animal breeding and genetics [plenary address]. Proc. World Congress on Genetics Applied to Livestock Production, Auckland, New Zealand. https://icarinterbullwcgalp.zerista.com/event/member/453201.

Howard, J., and S. Gugger. 2020. Fastai: A layered API for deep learning. Information (Basel) 11:108. https://doi.org/10.3390/info11020108.

Huang, G., Z. Liu, L. van der Maaten, and K. Q. Weinberger. 2017. Densely connected convolutional networks. Pages 2261–2269 in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (Institute of Electrical and Electronics Engineers), Piscataway, NJ.

Lainé, A., C. Bastin, C. Grelet, H. Hammami, F. G. Colinet, L. M. Dale, A. Gillon, J. Vandenplas, F. Dehareng, and N. Gengler. 2017. Assessing the effect of pregnancy stage on milk composition of dairy cows using mid-infrared spectra. J. Dairy Sci. 100:2863–2876. https://doi.org/10.3168/jds.2016-11736.

Lainé, A., H. Bel Mabrouk, L. Dale, C. Bastin, and N. Gengler. 2014. How to use mid-infrared spectral information from milk recording system to detect the pregnancy status of dairy cows. Commun. Agric. Appl. Biol. Sci. 79:33–38.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. Deep learning. Nature 521:436–444. https://doi.org/10.1038/nature14539.

McParland, S., G. Banos, E. Wall, M. P. Coffey, H. Soyeurt, R. F. Veerkamp, and D. P. Berry. 2011. The use of mid-infrared spectrometry to predict body energy status of Holstein cows. J. Dairy Sci. 94:3651–3661. https://doi.org/10.3168/jds.2010-3965.

Muhammd, F., A. Sarwar, and C. S. Hayat. 2000. Peripheral plasma progesterone concentration during early pregnancy in Holstein Friesian Cows. Pak. Vet. J. 20:166–168.

Olori, V. E., S. Brotherstone, W. G. Hill, and B. J. McGuirk. 1997. Effect of gestation stage on milk yield and composition in Holstein Friesian dairy cattle. Livest. Prod. Sci. 52:167–176. https://doi.org/10.1016/S0301-6226(97)00126-7.

Pan, S. J., and Q. Yang. 2010. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22:1345–1359. https://doi.org/10.1109/TKDE.2009.191.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12:2825–2830.

Penasa, M., M. De Marchi, and M. Cassandro. 2016. Short communication: Effects of pregnancy on milk yield, composition traits, and coagulation properties of Holstein cows. J. Dairy Sci. 99:4864–4869. https://doi.org/10.3168/jds.2015-10168.

Roelofs, J., F. López-Gatius, R. H. F. Hunter, F. J. C. M. van Eerdenburg, and C. Hanzen. 2010. When is a cow in estrus? Clinical and practical aspects. Theriogenology 74:327–344. https://doi.org/10.1016/j.theriogenology.2010.02.016.

Sheldon, M., and D. Noakes. 2002. Pregnancy diagnosis in cattle. In Pract. 24:310–317. https://doi.org/10.1136/inpract.24.6.310.

Shephard, R. W., and J. M. Morton. 2018. Estimation of sensitivity and specificity of pregnancy diagnosis using transrectal ultrasonography and ELISA for pregnancy-associated glycoprotein in dairy cows using a Bayesian latent class model. N. Z. Vet. J. 66:30–36. https://doi.org/10.1080/00480169.2017.1391723.

Smith, S. L., S. J. Denholm, M. P. Coffey, and E. Wall. 2019. Energy profiling of dairy cows from routine milk mid-infrared analysis. J. Dairy Sci. 102:11169–11179. https://doi.org/10.3168/jds.2018-16112.

Soyeurt, H., C. Bastin, F. G. Colinet, V. M. R. Arnould, D. P. Berry, E. Wall, F. Dehareng, H. N. Nguyen, P. Dardenne, J. Schefers, J. Vandenplas, K. Weigel, M. Coffey, L. Théron, J. Detilleux, E. Reding, N. Gengler, and S. McParland. 2012. Mid-infrared prediction of lactoferrin content in bovine milk: Potential indicator of mastitis. Animal 6:1830–1838. https://doi.org/10.1017/S1751731112000791.

Soyeurt, H., P. Dardenne, F. Dehareng, G. Lognay, D. Veselko, M. Marlier, C. Bertozzi, P. Mayeres, and N. Gengler. 2006. Estimating fatty acid content in cow milk using mid-infrared spectrometry. J. Dairy Sci. 89:3690–3695. https://doi.org/10.3168/jds.S0022-0302(06)72409-2.

Soyeurt, H., F. Dehareng, N. Gengler, S. McParland, E. Wall, D. P. Berry, M. P. Coffey, and P. Dardenne. 2011. Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries. J. Dairy Sci. 94:1657–1667. https://doi.org/10.3168/jds.2010-3408.

Toffanin, V., M. De Marchi, N. Lopez-Villalobos, and M. Cassandro. 2015. Effectiveness of mid-infrared spectroscopy for prediction of the contents of calcium and phosphorus, and titratable acidity of milk and their relationship with milk quality and coagulation properties. Int. Dairy J. 41:68–73. https://doi.org/10.1016/j.idairyj.2014.10.002.

Toledo-Alvarado, H., A. I. Vazquez, G. de los Campos, R. J. Tempelman, G. Bittante, and A. Cecchinato. 2018. Diagnosing pregnancy status using infrared spectra and milk composition in dairy cows. J. Dairy Sci. 101:2496–2505. https://doi.org/10.3168/jds.2017-13647.

van Rossum, G. 1995. Python tutorial, Technical Report CS-R9526. Centrum voor Wiskunde en Informatica (CWI), Amsterdam, the Netherlands. Accessed Jul. 10, 2018. https://docs.python.org/3/library/index.html.

Wojciechowski, K. L., and D. M. Barbano. 2016. Prediction of fatty acid chain length and unsaturation of milk fat by mid-infrared milk analysis1. J. Dairy Sci. 99:8561–8570. https://doi.org/10.3168/jds.2016-11248.

Yim, J., J. Ju, H. Jung, and J. Kim. 2015. Image classification using convolutional neural networks with multi-stage feature. Pages 587–594 in Advances in Intelligent Systems and Computing. J. Kacprzyk, ed. Springer Verlag, New York City, NY.

Zhuang, F., Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. 2019. A comprehensive survey on transfer learning. Proc. IEEE, Jan. 2021. 109:43–76. 10.1109/JPROC.2020.3004555.